Tracking Policy-relevant Narratives of Democratic Resilience at Scale: from experts and machines, to AI & the transformer revolution

Simon D Angus, Data & Policy, 2026

S1 Supporting Information

1 Traditional (pre-transformer) NLP approaches to automation

Whilst the standard approach to labelling text with frames (see main paper for details) is still widely used and accepted in rigorous analysis of public discourse, the recent digitisation of all areas of human social, economic and political life, and the massive quantity of discourse artefacts this has created (Mauro et al., 2016), has seen a move to automation of the standard approach, in particular, Step 5 – the coding of texts. Yet, as will be shown below, automating, and so scaling, the standard approach to framing analysis is just the starting point for automated methods with new technological frontiers opening up new ways of analysing and tracking public discourse.

Here, the fundamental challenge of computation lies at the heart of technological advance: how to encode the richness of human semantics in machine-readable format? Computational linguistics is the field of computer science perhaps most associated with attempting to answer this question, and multiple techniques have arisen over recent times which are still in use by computational social-, political-, and economic- scientists alike, to study public discourse (Bail, 2014; Gentzkow et al., 2019).

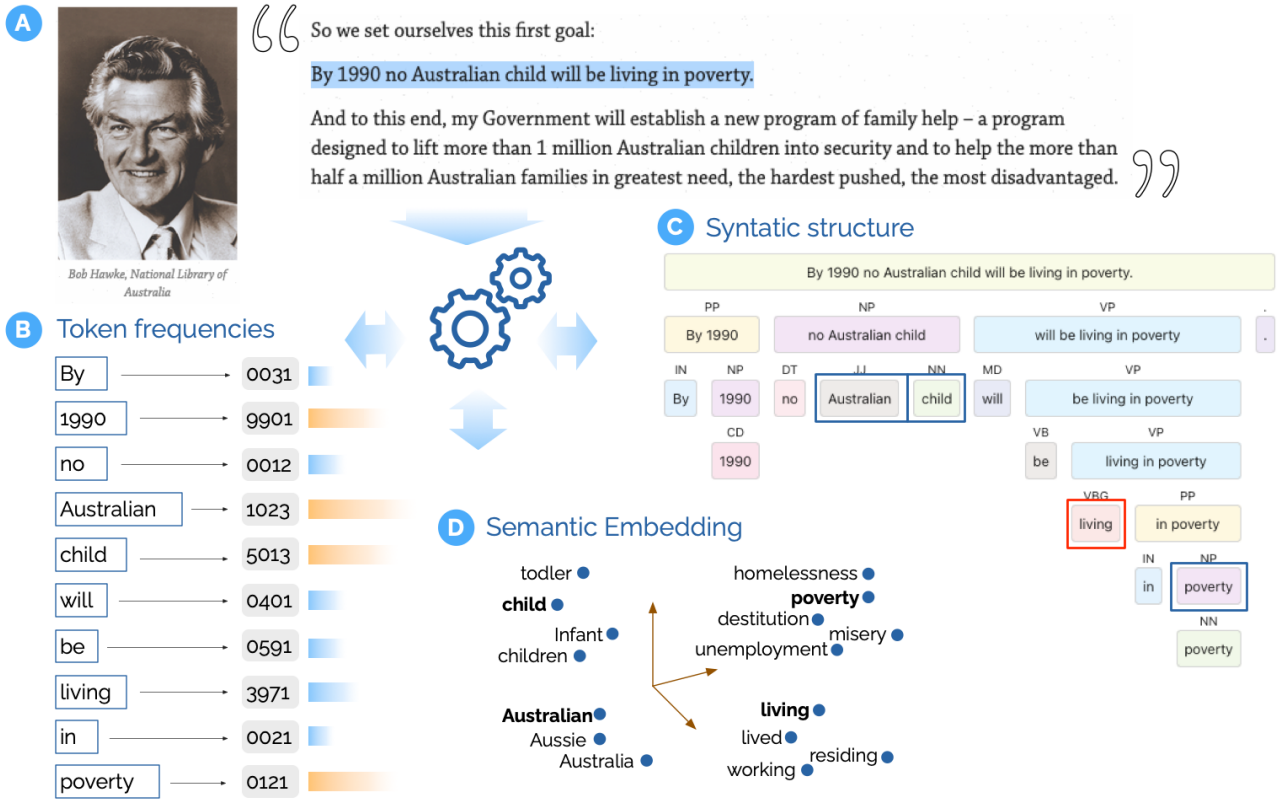

Figure A: **Traditional (pre-transformer) NLP approaches to analysing discourse.** Given some text (A), unique words (tokens) can be counted and turned into frequencies (B), syntactic structures can be identified and named-entities can be labelled (C), and words can be associated with their location in high-dimensional, pre-trained, semantic embedding spaces (D).

In figure **A** we summarise the fundamental technologies used in traditional (pre-transformer) NLP research applications in the social science. We discuss each in turn.

Token frequencies – A classic NLP approach is to define a unique vocabulary for a corpus of texts, typically dropping short, non-meaningful words - 'stop-words' ('a', 'to', 'is', 'the', etc.), and reducing words to their root or stem form (e.g. 'running' → 'run') so as to create a minimal set of meaningful word forms, known as 'tokens', by which any text can be represented. The computational task is then to count up the number of times each token appears in a given text ('token frequencies'). Importantly, such frequency representations lose any *contextual* information about the token. As such, these representations are called 'bag-of-word' approaches, since it is as if all of the words have been cut out of their context and tipped into a bag, regardless of order. Whilst such an approach may seem overly reductive, further processing can recover remarkable insights. For instance, by converting token frequencies ('tf') for a text into a vector, and then modifying token frequencies by a scarcity weighting for that token across all texts ('idf', inverse document frequency), researchers can then compute semantic distances between texts in this vector space.

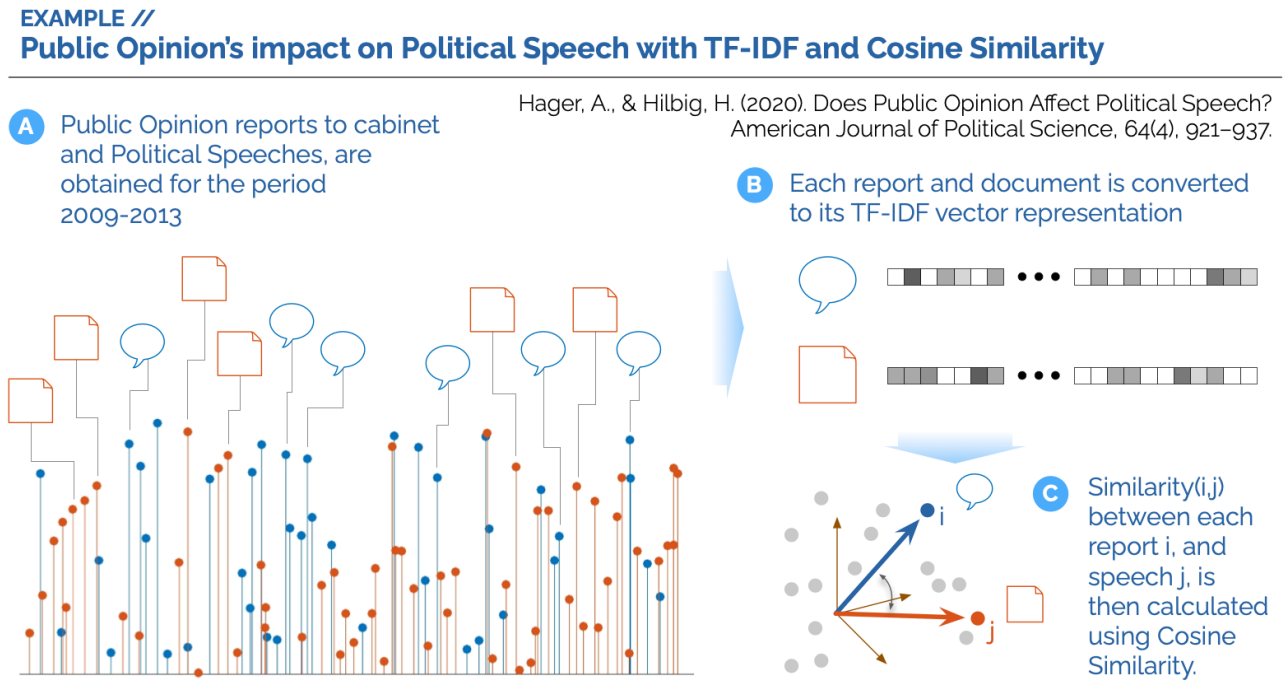

Figure B: **Causal analysis of public opinion's impact on political discourse.** In Hager and Hilbig (2020) a large corpus of German cabinet public opinion surveys were obtained after release, and, together with political speeches, press releases and other announcements of cabinet members (A), a regression discontinuity causal analysis framework was undertaken. After standard pre-processing, 3,860 length tf-idf vectors for each report and speech were calculated (B), enabling the calculation of semantic similarity between each report—speech pair via cosine similarity (C).

An example of such an approach is found in Hager and Hilbig (2020) who analyse public opinion survey texts along with political speeches to examine the causal relationship between public opinion and political discourse by calculating tf-idf vector distances (see Fig. **B**). Alternatively, tf-idf vectors can be fed into topic-modelling algorithms which perform unsupervised clustering[1] over the vectors typically using latent Dirichlet allocation (LDA) methods, such that topic dynamics can be followed over time (Jelodar et al., 2019). Examples of this strategy include Vidgen and Yasseri (2020b) who study the impact of petitioners on policy direction in the United Kingdom, Bonilla and Mo (2019) who analyse discourse topics in newspapers related to human trafficking, and Goyal and Howlett (2021) who study Covid-19 policy responses.

Other extensions to token frequency approaches include counting specific tokens which are said to encode attributes of language like sentiment. So-called 'dictionary-based' sentiment classifiers such as the Lexicoder Sentiment Dictionary (Young and Soroka, 2012) or VADER - Valence Aware Dictionary for Sentiment Reasoning (Hutto and Gilbert, 2014) are two examples of such approaches. Pauwels (2011) applies such an approach to analyse populism in Belgian party speeches and manifestos. The method introduced simply calculates the fraction of all words in a document appearing in the populism dictionary – effectively contending that populist

word usage is a tell-tale (i.e. *definition*) for populist narratives[2].

Whilst tf-idf methods are reasonably suited to topic modelling, the fact remains that since tf-idf vectors treat language use as entirely context-free, bag-of-word approaches are inherently limited in their ability to accurately identify nuanced semantic patterns which are inherent in narrative discourse analysis. Sentiment analysis being a good example of such limitations – dictionaries of terms associated with positive or negative sentiment are very labour-intensive to build, are domain specific, and are entirely context free, making sarcasm or irony a major stumbling block in their application (Grimmer and Stewart, 2013).

Semantic embeddings – A major revolution occurred in NLP at the start of the second decade of the millennia, with the development of massively trained neural word-embedding models. Here, neural network architectures are trained to guess the correct token from surrounding tokens, or alternatively, the surrounding tokens from a given token. For the first time, computational representation of human language went beyond mere frequencies of unique index values, to actually associating words with their meaning, as revealed by their (local) context of use. Such models are referred to as 'word embedding' models (or sometimes, simply, 'embeddings'), with Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and FastText (Bojanowski et al., 2017) among the most popular and widely used embeddings. By passing these models billions of sequences of human text (typically harvested from online sources), the models are able to accurately represent each word with a vector location in high-dimensional space. For instance, the vector for 'child' will be close - in vector space - to 'toddler', 'infant', and 'children'.

For the practitioner, word embeddings are very easy to accommodate, since any word (including stopwords; without stemming) can be passed to a pre-trained embedding (essentially a neural model), which deterministically produces a vector for that word. By repeating this process for a sequence of words in a text, many vectors are obtained, which can then be averaged to obtain a document location in 'semantic space'. From here, down-stream tasks like topic-modelling can be performed as introduced earlier.

However, there are more elaborate uses of word embeddings. A researcher can take the generically trained word-embedding model, and then further train the model on a sufficiently large sample of study texts, a process known as 'fine-tuning'. In this way, the word-embedding model itself starts to represent more closely the way that words encode semantics as revealed by the language in the study materials itself. An example of this approach is found in Kozlowski et al. (2019)'s ground-breaking study of class, where millions of book texts from each decade spanning the 1900s to the 1990s were each fed into a initialised Word2Vec model to fine-tune it to the particular semantic relationships of that decade. The authors then looked at semantic similarity between certain words which denote class and other words which denote dimensions of culture (see Fig. **C**). Whilst not directly assessing framing within narrative discourses, one can see how deep cultural associations – the underpinning of narrative framing (ala Polletta and Callahan (2017)) – can be quantified and tracked over time using such an approach.

Syntactic structure – Whilst grammar, on its own, does not closely associate with one narrative or frame, syntactic analysis empowers research methods to create more nuanced representations of narratives. Today, the field is almost entirely dominated by transformer methods, but there are still use-cases for traditional approaches, especially where data are scarce or computational resources are constrained. Here, hidden Markov Models or their derivatives represent pre-transformer approaches to labelling parts of speech McCallum et al. (2000); Brants (2000). Alternatively, other methods have been developed to correctly identify who is speaking in a sentence where pronouns replace names. This is especially important when conducting sentence-level analysis (a common approach in NLP), especially in political analysis, since knowing that the 'they' refers to the government, or opposition, or 'she' refers to a shadow-minister or the Prime Minister are critical pieces of information for any quantitative assessment. In NLP this task is referred to as 'coreference resolution' and fast, accurate methods are in common use (Clark and Manning, 2016a,0).

1.1 Ensemble approaches

In practice, traditional NLP workflows typically combine tools together in a sequence of steps, known as a 'pipeline' – a *combination* of transformations is used to achieve higher accuracy on a given research objective. This is the approach taken to a series of studies in computer science literature which define the task as a

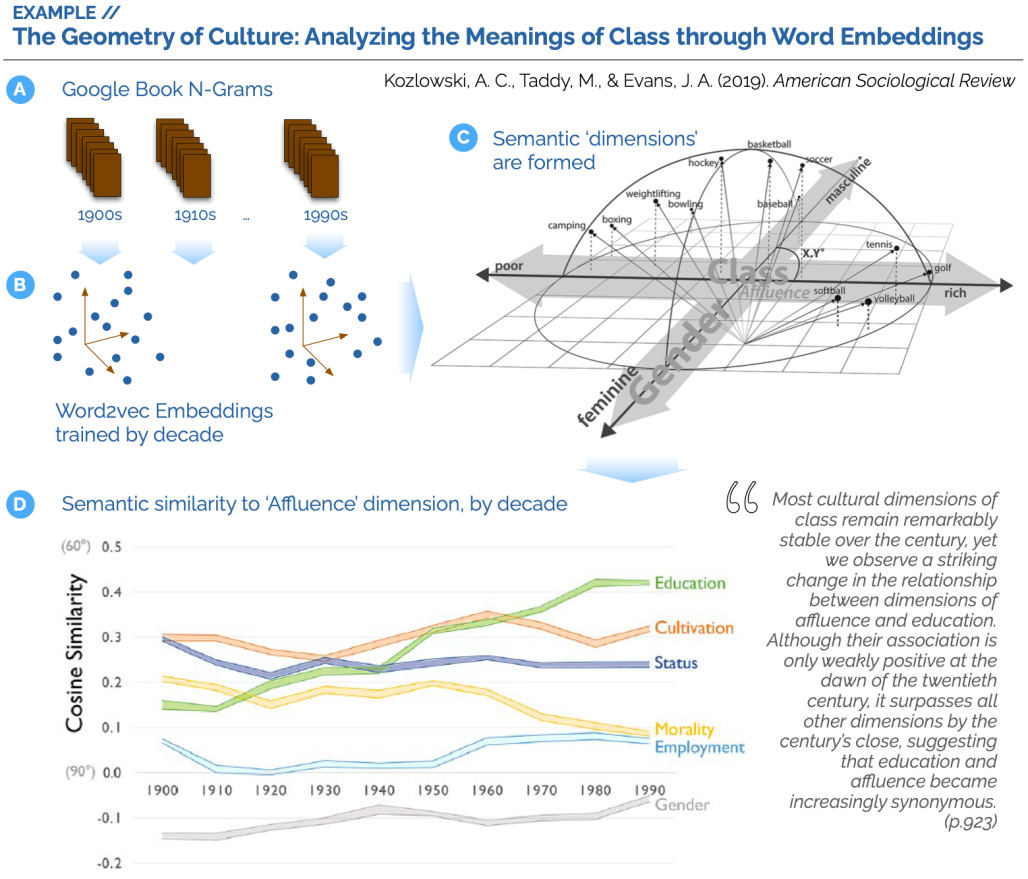

Figure C: **Using word-embeddings to analyse the foundations of narrative framing in millions of texts.** Kozlowski et al. (2019) take US provenance Google Books n-gram texts, per decade (A), to fine-tune Word2Vec neural embedding models (B), such that distances between semantic dimensions along word-pairs lines in each decadal model can be calculated (C), to track the cultural association of e.g. 'Affluence' with these dimensions over time (D).

multi-class classification 'framing' problem[3]. However, this literature is of limited application to tracking conceptual framing (in the spirit of Entman, i.e. narratives to *persuade* or *promote*) since 'framing' is used in this literature to denote a *dimension* of an *issue* rather than what might be called the conceptual framing itself. For example, Boydstun et al. (2013)'s 15 generic policy 'frames' include 'Economic', 'Public opinion', and 'Cultural identity' – these are generic dimensions of an issue and do not carry enough meaning to denote a specific narrative or conceptual framing. In the literature these 'frames' are then instrumented across multi-issue labelled datasets such as the Media Frames Corpus (MFC) (Card et al., 2015) and the Gun Violence Frame Corpus (GVFC) (Liu et al., 2019) and analysed using a variety of multi-class classification algorithms (Field et al., 2018; Zhang et al., 2023).

However, not all computational literature follows this path, with several studies taking the more nuanced approach to framing analysis required for tracking narratives of democratic resilience. For example, Vidgen and Yasseri (2020a) offer a promising approach to tracking nuances in Islamic hate-speech in 4,000 social media posts. Namely, they distinguish between 'strong' and 'weak' islamophobia – a distinction which can be considered as two frames within the general attitude of islamophobia. In the study, they trial seven different pipelines, leveraging all of the traditional NLP approaches mentioned thus-far: tf-idf vectors, word embeddings, fine-tuned word-embeddings, sentiment, and syntactic features. They also trial a number of different classification algorithms from logistic regression, Naive-Bayes, through random forests, and support vector machines (SVM) and even deep learning (neural methods). The most accurate approach they identify combines features, and uses the SVM classifier, obtaining 72.17% accuracy on their training data ($n = 1,341$), and 77.3% accuracy on a further set of 300 hand-labelled test tweets. Likewise, Morstatter et al. (2018) study narratives of support and opposition (polarity) across 10 framings related to the balistic missile defence (BMD) issue in Europe. However, through hand-coding, and with training a multi-class NLP classifier, accuracy is low (less than 0.5 across classes). Cocco and Monechi (2022) pursue populism narrative analysis via adding machine learning to a bag-of-words approach, processing over 240,000 sentences from 268 electoral manifestos across 99 parties using a random forest (RF) classifier. Here, rather than a list of populist words, the method mechanically labels (i.e. *defines*) as 'populist' all sentences that arise from texts of parties labelled populist in the PopuList database (Rooduijn et al., 2024). The RF tool is trained to pick sentences from these texts (represented by cleaned and stemmed word occurrence), as opposed to non-populist texts in the training data.

Closer still to the narrative analysis task are studies which seek to track the incidence of 'micro-narrative' phrases within a large body of texts. Here, the work of Ash et al. (2024) and Anantharama et al. (2022) seek to identify 'Entity – Verb – Entity' (EVE) triplets, through a multi-step pipeline using word-embeddings (Ash et al.) or transformer-based sentences embeddings (Anantharama et al.), together with syntactic parsing and co-reference resolution. These approaches enable the tracking of phrases like 'God–bless–America' or 'indigenous voice–enshrine–constitution'. By clustering a set of these micro-narratives and associating with a particular framing, narrative components can be tracked across speakers, and through time.

So how well do non-transformer methods perform at traditionally hand-coded approaches to tracking discourse? A study by Nelson et al. (2021) considers three traditional methods (dictionary-based, tf-idf supervised, unsupervised topic discovery) applied to a large number of hand-coded news articles with a hierarchical identification schema relating to inequality. They find that tf-idf style classifiers were the most accurate approach in both identifying articles and tracking themes over time. However, with a combined precision/recall (f1) score of around 0.8, it would be hard to conclude that 'human-level accuracy' was obtained by the best method. However, the authors took methods 'off the shelf' and it is likely that more performance could be obtained by skilled practitioners.[4]

References

Anantharama, N., Angus, S. D., and O'Neill, L. (2022). Canarex: Contextually aware narrative extraction for semantically rich text-as-data applications. *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3551–3564.

Ash, E., Gauthier, G., and Widmer, P. (2024). Relatio: Text semantics capture political and economic narratives. *Political Analysis*, 32:115–132.

Bail, C. A. (2014). The cultural environment: Measuring culture with big data. *Theory and Society*, 43:465–482.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bonilla, T. and Mo, C. H. (2019). The evolution of human trafficking messaging in the united states and its effect on public opinion. *Journal of Public Policy*, 39:201–234.

Boydstun, A. E., Gross, J. H., Resnik, P., and Smith, N. A. (2013). Identifying media frames and frame dynamics within and across policy issues. In *Proceedings of New Directions in Analyzing Text as Data Workshop*. Part of the New Directions in Analyzing Text as Data Workshop.

Brants, T. (2000). TnT – a statistical part-of-speech tagger. In *Sixth Applied Natural Language Processing Conference*, pages 224–231, Seattle, Washington, USA. Association for Computational Linguistics.

Card, D., Boydstun, A. E., Gross, J. H., Resnik, P., and Smith, N. A. (2015). The media frames corpus: Annotations of frames across issues. In Zong, C. and Strube, M., editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Clark, K. and Manning, C. D. (2016a). Deep reinforcement learning for mention-ranking coreference models. *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 2256–2262.

Clark, K. and Manning, C. D. (2016b). Improving coreference resolution by learning entity-level distributed representations. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, 2:643–653.

Cocco, J. D. and Monechi, B. (2022). How populist are parties? measuring degrees of populism in party manifestos using supervised machine learning. *Political Analysis*, 30:311–327.

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43:51–58.

Field, A., Kliger, D., Wintner, S., Pan, J., Jurafsky, D., and Tsvetkov, Y. (2018). Framing and agenda-setting in russian news: a computational analysis of intricate political strategies. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 3570–3580.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57:535–74.

Goyal, N. and Howlett, M. (2021). "Measuring the Mix" of Policy Responses to COVID-19: Comparative Policy Analysis Using Topic Modelling. *Journal of Comparative Policy Analysis: Research and Practice*, 23(2):250–261. Publisher: Routledge.

Grimmer, J. and Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3):267–297. Edition: 2017/01/04 Publisher: Cambridge University Press.

Hager, A. and Hilbig, H. (2020). Does public opinion affect political speech? *American Journal of Political Science*, 64:921–937.

Hutto, C. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8. Issue: 1.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., and Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211.

Kozlowski, A. C., Taddy, M., and Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84:905–949.

Liu, S., Guo, L., Mays, K., Betke, M., and Wijaya, D. T. (2019). Detecting frames in news headlines and its application to analyzing news framing trends surrounding u.s. gun violence. *CoNLL 2019 - 23rd Conference on Computational Natural Language Learning, Proceedings of the Conference*, pages 504–514.

Mauro, A. D., Greco, M., and Grimaldi, M. (2016). A formal definition of big data based on its essential features. *Library Review*, 65:122–135.

McCallum, A., Freitag, D., and Pereira, F. C. N. (2000). Maximum entropy markov models for information extraction and segmentation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, page 591–598, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*.

Morstatter, F., Wu, L., Yavanoglu, U., Corman, S. R., and Liu, H. (2018). Identifying framing bias in online news. *ACM Transactions on Social Computing*, 1:1–18.

Nelson, L. K., Burk, D., Knudsen, M., and McCall, L. (2021). The Future of Coding: A Comparison of Hand-Coding and Three Types of Computer-Assisted Text Analysis Methods. *Sociological Methods & Research*, 50(1):202–237.

Pauwels, T. (2011). Measuring populism: A quantitative text analysis of party literature in belgium. *Journal of Elections, Public Opinion and Parties*, 21:97–119.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Polletta, F. and Callahan, J. (2017). Deep stories, nostalgia narratives, and fake news: Storytelling in the trump era. *American Journal of Cultural Sociology*, 5:392–408.

Rooduijn, M., Pirro, A. L., Halikiopoulou, D., Froio, C., Kessel, S. V., Lange, S. L. D., Mudde, C., and Taggart, P. (2024). The populist: A database of populist, far-left, and far-right parties using expert-informed qualitative comparative classification (eiqcc). *British Journal of Political Science*, 54:969–978.

Vidgen, B. and Yasseri, T. (2020a). Detecting weak and strong islamophobic hate speech on social media. *Journal of Information Technology & Politics*, 17:66–78.

Vidgen, B. and Yasseri, T. (2020b). What, when and where of petitions submitted to the UK government during a time of chaos. *Policy Sciences*, 53(3):535–557.

Young, L. and Soroka, S. (2012). Affective News: The Automated Coding of Sentiment in Political Texts. *Political Communication*, 29(2):205–231. Publisher: Routledge.

Zhang, Y., Shah, D., Pevehouse, J., and Valenzuela, S. (2023). Reactive and asymmetric communication flows: Social media discourse and partisan news framing in the wake of mass shootings. *International Journal of Press/Politics*, 28:837–861.